Data Workflow

Data Management Workshop

ELIZABETH WICKES, DATA CURATOR
HEIDI IMKER, DIRECTOR
RESEARCH DATA SERVICE
Research Data Service (RDS)

The Research Data Service provides the Illinois research community with expertise, tools, and infrastructure to manage and steward research data.

- Knowledge around data policies, resources, archiving, & preservation
- Consultation for data management planning & implementation
- Workshops on data management, documentation, and data publishing
- Data Management Plan reviews and DOI minting services
- Solutions for public access to research data
- Centralized, private storage for active (“working”) data (with NCSA)

visit: researchdataservice.illinois.edu or email: researchdata@library.illinois.edu
What do we do?

**Expertise**
- Knowledge around data policies, tools, resources, archiving, and preservation
- Consultation and workshops for data management planning and implementation

**Tools**
- Data Management Plan creation wizard (DMPTool.org)
- Tools for data citation (DOI minting)

**Infrastructure (in progress)**
- Illinois Data Bank (self-deposit institutional data repository)
Workflow Workshop Goals

• Know
  • the tools you use
  • the data you use
  • where it all lives
  • where it all goes

• Learn
  • How your data workflow works
  • Points where you need clarification
  • How your collaboration with others could be improved

• Practice
  • Mapping out your workflow
Required Materials

• 1 color, 3x3 post it note
• 4 colors, smaller post it notes
  • 1 color the same as the 3x3 note
• Large paper, divided into 3 horizontal sections
• Handout with cheat sheet
• Pen or pencil
Why is documentation important (in the short term)?

• Because other humans exist
• Those humans will need to make use of your stuff
• Many of those humans are sitting next to you right now
• But humans move around, particularly students
• So at some point, there will be humans in this room, using your stuff, and have no idea who you are.
Like retracing your steps...

• When you lose your phone or keys, one of the best ways to figure out where it might be is retracing your steps.

• You think about what you’ve done, step by step, and in theory you can recreate everything you’ve touched in the process.

• Thus, workflow mapping can be a useful first step in identifying where you may have tucked data away.
What data do you have?

Input
- Source data
- Data from other people

Process
- Temporary files
- Intermediate datasets

Output
- Output data
- Data for other people
- Data that goes into reports or other final products
What do you do to that data?

- Input
  - Ingest

- Process
  - Clean
  - Train
  - Test

- Output
  - Analysis
  - Write up
  - Backup
So how do you science?

Investigate data

Input data

Input data

Input data

Get other data in

Clean the data

Make test data

Train a model

Test the model

Make some charts

Check the algorithm

Write scripts

Save stats analysis

Clean the data again

Join in other data

Output data

Output data

Output data
So how do you science?

- Input data
- Input data
- Input data

- Make some
- input data
- Investigate
- input data
- Join in
- input data

- Make some
- test data
- Make test
- data

- Train a
- model
- Test the
- model

- Get other
- data
- Check the
- algorithm

- Write some
- scripts
- Make some
- charts
- Join in
- other data

- Save stats
- Analysis
- Clean the
- data again

- Output data
- Output data
- Output data

SCIENCE.
So how do you science?

**Input data**
- Input data
- Input data
- Input data

**Output data**
- Output data
- Output data
- Output data

**Inves<gate**
- Inves<gate
- Inves<gate
- Inves<gate

**Test**
- Test
- Test
- Test

**Clean**
- Clean
- Clean
- Clean

**Make**
- Make
- Make
- Make

**Train**
- Train
- Train
- Train

**Test**
- Test
- Test
- Test

**Get**
- Get
- Get
- Get

**Join**
- Join
- Join
- Join

**Save**
- Save
- Save
- Save

**Publications**
- Don’t forget about us!

**SCIENCE.**
So you’ve got stuff

• When working in large teams, it is particularly important to know:
  • What you receive from others (input data and materials)
  • What you make for others
Activity: Workflow Map

• This will be our main activity today, so feel free to take your time on these steps and ask questions.

• The intention is not to capture every detail of your workflow, but to help you get a feel for the big picture and points where you may need clarification or other help.
Example workflows
Example workflows

1. **Input Surveys by hand**
   - Continue until over 90% of facilities have reported

2. **Check data with Excel tools**
   - Input corrected data

3. **Contact facilities with unusual data or incorrectly marked forms**

4. **LMO Report Data**

5. **Track the LMO allocation**

6. **Query the db for annual report totals**

7. **Spreadsheet on the Server**

8. **Queries on the Server**

9. **Ad hoc research requests**

10. **Grid modeling data**

11. **Annual report data**
The Board

Inputs

- Data Products
- Activities
- Scripts, Software, & Tools

Outputs

- Data Input/Output (from others)
- Data (NOT for others)
- Tools Used
- Notes or Annotations

Workflow activities
Make this your own

• You know what you do best
• Use your own voice and words
• Just be sure you’ll be able to understand them later
Step 1: Identify your inputs

• Take the pink post-its and write “input” in the top left corner

• Think of the project you want to map in this workflow. Write the input data for this project on the pink post-it.
  • This can be raw data
  • This can be data you get from someone else, and write down their name as well!
  • If you get data from more than one source, you can make more than one post-it. Just be sure to label them all “input.”

• Put the pink post-it to the left side of the center section of your sheet of paper.
Data Input (from others)

Data Input (from others)
Step 1: Start working on the steps

• Take 4 regular yellow post-its and line them up in the middle of your paper, after the input post-it. Add more as necessary, but try to think at a very high level.

• Use the smaller yellow post-its to add annotations.

• Examples:
  • ingest data
  • make training set
  • train model
  • test model
  • etc.
Data Input (from others)

Activity 1...

Activity 2...

Activity 3...

Activity 4...

Loop here until model fits
Step 2: How do you process the data?

- Look at the individual steps you’ve written down:
  - Is there a script, software package, reference resource, or other tool you need in order to complete the step? If so, write it on one of the **darker blue post-its**.
  - Are there temporary, scratch, or other data files created for during this step? If so, make note of them on a **lighter blue post-it**
  - Are you making data for other people in this step? Use a **pink post-it** Write down their name(s) on it as well.

- Continue this process with each of the steps of your workflow
Data Input (from others)

Activity 1...
Activity 2...
Activity 3...
Activity 4...

Loop here until model fits

script1.py
some other software

test dataset
training dataset
model
sqlite database
visualizations for paper

Data Input (from others)

test dataset
training dataset
model
sqlite database
visualizations for paper

script1.py
some other software
analysis.r

loop here until model fits
Step 3: What are my outputs?

• When you get to steps where you create data that is handed off to other people or that you need someone else’s help to complete, note that information down on a pink post-it and label it “output.”

• Put output post-its at the end of your workflow line, or above the step where they’re generated if they’re produced before your workflow is complete.
Data Input (from others)

Activity 1...

Activity 2...

Activity 3...

Activity 4...

Loop here until model fits

script1.py

some other software

sqlite database

model

test dataset

training dataset

report for developer

model fitting script for dev

visualizations for paper

report for developer

some other software

analysis.r
Step 4: Where were there problems?

- Did you run into something you don’t know, need to look up, or need to finish? Make a note on a red post-it and place it by the appropriate point in your workflow.
Activity discussion

• What did we learn from this?

• Are there points where you need more interaction with your team than you realized?

• What data is for your personal use only? Where is it stored and how do you manage it?

• What data needs to go to other people? How are you sharing it? Are you keeping backups of it as well? Where?

• Homework: Take a picture of this workflow back to your team. Where do their workflows hook into yours? Where do yours hook into theirs? Are there ways you can improve data sharing within the team?
Why is documentation important (for the long term)?
National Data Policy

OSTP MEMO: INCREASING ACCESS TO THE RESULTS OF FEDERALLY FUNDED SCIENTIFIC RESEARCH

“requiring researchers to better account for and manage the digital data resulting from federally funded scientific research”

• Data management plans will be come compulsory

• Providing public access to data will become more routine

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research
Publisher Policies

Science

GENOME RESEARCH

nature.com

PLOS

Bioinformatics

Proceedings of the National Academy of Sciences of the United States of America
Publisher Policies

• *Bioinformatics*

“All data on which the conclusions given in the publication are based must be publicly available...”

• *Genome Research*

“*Genome Research* will not publish manuscripts where data used and/or reported in the paper is not freely available in either a public database or on the *Genome Research* website. There are no exceptions.”
(Discussions of) Mistakes Are Public
Data "Publication"

Making datasets (in and of themselves) publically accessible

• improves transparency and reproducibility of research
• save time by reducing duplication of effort (yours or theirs)
• makes the data itself independently discoverable
• another way to expose to your work
• maybe you’ll need that data again some day
In Action – Warnow Paper

In Action – Warnow Data


Downloads as of Sept 26th: 1522
Illinois Data Bank (databank.illinois.edu)

A self-serve publishing platform that centralizes, preserves, and provides persistent and reliable access to Illinois research.

- can be linked to related materials, such as articles, theses, code, and other datasets
- can include files of any format and sizes up to 15 GB/file via Box.com
- can be deposited for immediate release or temporarily embargoed
- receive a stable, unique identifier (DOI) for persistent access and ease of citation
- are registered in a central, world-wide catalog for better discovery
- are professionally managed and curated by the Research Data Service staff at the University Library
- are preserved for a minimum of 5 years